Are there gender differences in cognitive reflection? Invariance and differences related to mathematics.

Primi, C., Donati, M. A., Chiesi, F., & Morsanyi, K. (2017). Are there gender differences in cognitive reflection? Invariance and differences related to mathematics. *Thinking & Reasoning*, 1-22. https://doi.org/10.1080/13546783.2017.1387606

Published in:
Thinking & Reasoning

Document Version:
Peer reviewed version

Are there gender differences in cognitive reflection?

Invariance and differences related to mathematics

Caterina Primi*, Maria Anna Donati*, Francesca Chiesi* & Kinga Morsanyi°

*NEUROFARBA – Section of Psychology, University of Florence (Italy)
°School of Psychology, Queen's University Belfast (UK)

Corresponding author:
Caterina Primi
Neurofarba – Section of Psychology
University of Florence (Italy)
Via S.Salvi 12 – Padiglione 26
50135 Florence - Italy
primi@unifi.it

Please cite paper as:

Primi, C., Donati, M., Chiesi, F. & Morsanyi, K. (2018). Are there gender differences in cognitive reflection? Invariance and differences related to mathematics. *Thinking & Reasoning*. DOI: 10.1080/13546783.2017.1387606

Abstract: Cognitive reflection is recognized as an important skill, which is necessary for making advantageous decisions. Even though gender differences in the Cognitive Reflection test (CRT) appear to be robust across multiple studies, little research has examined the source of the gender gap in performance. As a preliminarily step, we performed a meta-analysis of studies conducted by our research group in recent years to investigate the size of gender differences in cognitive reflection. Then, in Study 1, we tested the invariance of the scale across genders. In Study 2, we investigated the role of math anxiety, mathematical reasoning, and gender in CRT performance. The results attested the measurement equivalence of the *Cognitive Reflection Test – Long* (CRT- L), when administered to male and female students. Additionally, , the results of the mediation analysis showed an indirect effect of gender on CRT-L performance through mathematical reasoning and math anxiety. The direct effect of gender was no longer statistically significant after accounting for the other variables. The current findings suggest that cognitive reflection is affected by numerical skills and related feelings, such as math anxiety, which, in turn, explains the gender differences.

Cognitive reflection is recognized as an important skill, which is necessary for making advantageous decisions. Even though gender differences in the Cognitive Reflection test (CRT) appear to be robust across multiple studies, little research has examined the source of the gender gap in performance. In Study 1, we tested the invariance of the scale across genders. In Study 2, we investigated the role of math anxiety, mathematical reasoning, and gender in CRT performance. The results attested the measurement equivalence of the *Cognitive Reflection Test – Long* (CRT- L), when administered to male and female students. Additionally, the results of the mediation analysis showed an indirect effect of gender on CRT-L performance through mathematical reasoning and math anxiety. The direct effect of gender was no longer statistically significant after accounting for the other variables. The current findings suggest that cognitive reflection is affected by numerical skills and related feelings.

Keywords: Cognitive Reflection; Gender Differences; Invariance; Item Response Theory; Math Anxiety; Mathematical Reasoning; Mediation Model, Meta-Analysis.

Introduction

The Cognitive Reflection Test (CRT; Frederick, 2005) is a hugely popular measure of the tendency to avoid errors based on intuitive response tendencies, and to rely on careful deliberation. As an example, consider the following item: *A bat and a ball cost $1.10 in total. The bat costs $1.00 more than the ball. How much does the ball cost? _____ cents.* Although the correct response is 5 cents, many participants give the response "10 cents", which seems to pop into mind effortlessly. Indeed, a remarkable property of the CRT is that for each item, almost all participants produce either the normatively correct response, or a typical incorrect (i.e., heuristic) response. Cognitive reflection involves the ability to effectively monitor and correct impulsive response tendencies, and it is related to a wide variety of cognitive and decision-making skills (e.g., Cokely & Kelley, 2009; Frederick, 2005; Koehler & James, 2010; Oechssler, Roider, & Schmitz, 2009; Toplak, West & Stanovich, 2011; 2014), as well as thinking dispositions (Fernbach, Rogers, Fox, & Sloman, 2013; Fernbach, Sloman, Louis, & Shube, 2013; Mata, Fiedler, Ferreira, & Almeida, 2013; Pennycook, Cheyne, Koehler, & Fugelsang, 2016; Shenhav et al., 2012).

Frederick (2005), in his original report on the Cognitive Reflection Test (CRT), found that men outperformed women, and gender differences in the CRT have been confirmed in several subsequent studies with participants from different age groups, educational levels, and countries, and using the original CRT as well as modified versions of the original test. These results are surprising, because Frederick (2005, p.26) described the CRT as a measure of "one type of cognitive ability". Given that gender differences in cognitive abilities are not commonly found, these results raise the important question of whether the CRT measures the same trait in men and women (i.e., whether it is an appropriate measure of reflectivity in the case of both genders). In case the underlying trait measured by the test is the same for both genders, an additional question is whether the gender difference can be explained by any particular cognitive or affective factor that is related to performance on the CRT. The aim of the current paper was to address these questions.

Studies with adults (Campitelli & Gerrans, 2014; Cueva et. al., 2016; Pennycook, Cheyne, Koehler, & Fugelsang, 2016) found that males scored higher on the CRT than females, and females gave more intuitive responses than males, while no gender differences emerged for other types of incorrect answers that did not correspond to the typical intuitive response. Campitelli and Gerrans (2014) also showed that women struggled with inhibiting the intuitive response, especially in the case of the "bat and ball" problem. An analogous relation between gender and performance on the original CRT was found by Sinayev and Peters (2015) with American adults, by Ring, Neyse, David-Barett and Schmidt (2016), who tested German undergraduate students, and by Albaity, Rahman and Shahidul (2014) whose study involved Malaysian adults from different ethnic groups.

Gender differences were also found in the case of extended versions of the CRT. In a study conducted by Toplak, West and Stanovich (2014) with Canadian university students, male students obtained higher scores than female students, not only on the original CRT, but also on four new items. Primi, Morsanyi, Chiesi, Donati and Hamilton (2016) confirmed that males outperformed females on the original CRT, and also observed a gender difference in the case of a long form of the CRT (CRT-L), which included three new items. Their results were obtained with Italian and British students, attending the senior year of high school and undergraduate university courses. In sum, the studies that investigated gender differences on the CRT have found that males perform better than females on every single question, and that females are more likely to answer none of the questions correctly (i.e., they are more likely to show very low levels of cognitive reflection), while males are more likely to answer all three questions correctly (i.e., to exhibit very high levels of cognitive reflection). Importantly, gender differences persist even when controlling for test characteristics (e.g., monetary incentives, computerized administration, student samples, and positioning of the experiment; see Brañas-Garza, Kujal & Lenkei, 2015, for a review of 118 cognitive reflection test studies).

In order to obtain an estimate of the effect size of gender differences in cognitive reflection, we have performed a meta-analysis of studies conducted by our research group in recent years

(Morsanyi, Primi, Handley, Chiesi & Galli, 2012; Morsanyi et al., 2014; Morsanyi, McCormack & O'Mahony, 2017; Primi et al., 2016).[1] Although in some of these studies we used the CRT-Long, we have only considered performance on the original CRT items, so that we could combine scores from a larger number of studies.[2] The meta-analysis included data from 2,536 participants (1,611 females) from 13 samples. Most participants were undergraduate students from the UK and Italy. Some studies also included adolescent participants, and in one study a multiple-choice version of the CRT was administered. These details are listed in Figure 1. The analysis was conducted using the Comprehensive Meta-Analysis software using a random-effects model (i.e., we assumed that the true effect size of gender differences might vary from study to study). Specifically, we expected that there might be a variation due to the heterogeneity of our samples in terms of age and level of education. The summary effect for the meta-analysis appears in the last row of Figure 1.[3] The point estimate was .53 (*SE*=.10, lower limit =.34, upper limit = .72), which corresponds to a medium effect size for gender differences in cognitive reflection.

Even though gender differences in the CRT appear across multiple studies, only a few studies have examined the source of the gender gap in performance. As the CRT items have mathematical content, several studies have investigated the relationship between cognitive reflection and mathematical ability. Research has shown that people who perform well on the CRT tend to perform well on numeracy tests. Specifically, significant, moderate correlations have been found between performance on the CRT and mathematical ability, as indexed by numeracy, math achievement, or math skills (*r*=.25, Gómez-Chacón, García-Madruga, Vila, Elosúa, & Rodríguez, 2014; *r*=.31, Cokely & Kelley, 2009; Primi et al., 2016; *r*s ranging from .37 to .51, Liberali et al.,

[1] This analysis was specifically conducted for the purposes of the current paper, and has not been previously reported elsewhere. We have also included the data from the current paper in this analysis.

[2] In studies where we used the CRT-Long, the original items were administered before the new items. For this reason, performance on the original CRT items was not affected by the inclusion of the new items.

[3] Under the random-effects model, there is a distribution of true effects, and the summary effect is an estimate of the mean of this distribution.

2012; *r*=.43, Campitelli & Gerrans, 2014; Weller, Dieckmann, Tusler, Mertz, Burns, & Peters, 2013; *r*=.44 with numeracy and *r*=.61 with calculation, Sinayev & Peters, 2015; *r*s ranging from .29 to .45, Morsanyi, Busdraghi, & Primi, 2014; *r*=.53, Finucane & Gullion, 2010; *r*=.58, Thomson & Oppenheimer, 2016).

Some authors suggested that the numerical content of the CRT items might be responsible for the gender differences. For instance, Primi et al. (2016) found that the gender difference in the CRT-L was significantly reduced after the effects of numeracy were controlled, and they became non-significant when the effect of subjective numeracy was controlled. Zhang et al. (2016) verified that gender differences could be entirely explained by differences in subjective numeracy, that is, individuals' perceived competence in dealing with quantitative information, (in other words, their self-efficacy in the quantitative domain). Consistent with these findings, Thomson and Oppenheimer (2016) found that men outperformed females in the original CRT, but they did not find a gender difference in their newly developed four items that did not require any numerical computations (although they did involve numbers).

Starting from the premise that the CRT has a significant math component, the aim of the current study was to investigate the possibility that gender differences were related to the numerical content of the problems. Indeed, several studies have reported gender differences in math ability (e.g., Benbow, Lubinski, Shea, & Eftekhari-Sanjani, 2000; Hyde, Fennema & Lamon, 1990; Mau & Lynn, 2000), showing that males outperformed females. Probably the strongest evidence for the existence of gender differences in mathematics performance comes from the Programme for International Student Assessment (PISA), published by the Organization for Economic Co-operation and Development (OECD, 2016), which assesses the competencies of 15-year-old students from 65 different countries in various subjects, including mathematics. On average, across the OECD countries, boys outperform girls in mathematics by eight score points[4]. Between the PISA 2012 and PISA 2015 assessments, the gender gap did not change significantly in the vast majority of countries. Nevertheless, both the existence and the

[4] The same report also shows a relatively large gender gap in reading in favour of girls, and negligible gender differences in science performance.

nature of gender differences have been questioned. For example, Hyde, Lindberg, Linn, Ellis and Williams (2008), using data from over 7 million students, found no evidence of gender differences on US state math tests among students between grade 2 and grade 11.

Cognitive reflection has not only been found to be related to math skills, but also to the self-assessment of feelings and perceptions concerning one's ability to reason about and solve mathematical tasks. For instance, performance on the CRT was found to be significantly and positively related to participants' subjective perception of their quantitative abilities (r=.19, Primi et al., 2016; r=.39, Zhang et al., 2016; ranging from .43 to .48, Liberali et al., 2012), and it was significantly and negatively related to math anxiety (MA) (r=-.20, Morsanyi et al., 2014). MA has also been investigated across PISA countries in the 2012 survey. The results showed that around 30% of students reported feeling helpless or nervous when faced with math problems, and negative feelings toward math are also associated with lower school performance. Special attention has been paid to math anxiety (MA) and its impact on mathematical learning: an ever-growing body of research has recognized that anxiety states and worry experienced during math classes or related activities are significant factors with a negative influence on math learning and basic numerical abilities in both adults (Bursal & Paznokas, 2006; Jameson & Fusco, 2014; Pozehl, 1996; Maloney & Beilock, 2012; McMullan, Jones, & Lea, 2010; Swars, Daane, & Giesen, 2006) and children (Hill, et al., 2016; Wu, Barth, Amin, Malcarne, & Menon, 2012). MA thus seems to have serious consequences, not only in the short term (on math performance at school), but also in the long term, adversely influencing an individual's choice of career, type of occupation, and professional growth in adulthood (Ashcraft & Ridley, 2005; Beasley, Long & Natali, 2001; Chimman, Krantz, & Silver, 1992; Hembree, 1990; Ho et al., 2000). Concerning gender differences, the findings generally suggest that females suffer from MA more than males (see Else-Quest, Hyde & Linn, 2010; and see Devine, Fawcett, Szucs, & Dowker, 2012, for a short review), and that women are consequently less likely to seek opportunities for math problem solving, and they tend to avoid math-related

activities (Baloğlu & Kocak, 2006; Else-Quest et al., 2010; Jain & Dowson, 2009; McGraw, Lubienski, & Strutchens, 2006; Rubinsten, Bialik, & Solar, 2012).

Starting from these previous studies, we wanted to investigate more thoroughly the origin of the gender differences in cognitive reflection. Nevertheless, before delving into this issue, it is necessary to check the measurement invariance of cognitive reflection across genders. If a test is not invariant (i.e., if it does not measure the same construct in the same way in different groups of respondents), the comparison of test scores between different groups of individuals has to be considered invalid (Waiyavutti, Johnson, & Deary, 2012).

To the best of our knowledge, so far the invariance of the cognitive reflection test across genders has never been tested. That is, it is not clear if the items of the test are suitable to measure the construct (cognitive reflection) in males as well as females. The analysis of Differential Item Functioning is central to the investigation of the measurement equivalence of a scale at the item level (i.e., DIF allows one to assess whether the items measure the same trait dimension when administered to two different groups). Thus, it is necessaryto verify that the structure of the test, and item functioning is the same amongst male and female samples, in order to ascertain that the documented gender differences are due to real differences among males and females and not the result of biases in item functioning.

In sum, the aim of Study 1 was to provide evidence of the gender invariance of the *Cognitive Reflection Test – Long* (CRT-L), a new version of the CRT, which was recently developed in order to obtain a valid and reliable instrument, which overcomes some of the limitations of the original three-item CRT (Primi et al., 2016). If the CRT-L is invariant across genders, observed scores depend only on the levels of the latent construct, and not on group membership, and observed differences between groups reflect true differences in the amount or variability of the construct. With these premises, Study 2 sought to investigate the possibility that gender differences were related to the numerical content of the problems. In particular, we tested the hypothesis that math

skills and math anxiety explained gender differences in cognitive reflection.

Study 1

When comparing groups, researchers assume that the instrument (questionnaire, ability test) that they employ measures the same construct in all groups. Despite its appeal and its practical significance, this assumption is often not justified and needs to be tested. If the test does not measure the same construct across different groups, results are not comparable and inferences about group differences are misleading. The general term used to describe the lack of correspondence between measures applied to different groups is bias (Van de Vijver & Poortinga, 1997). Measurement equivalence might be threatened by different forms of bias. *Item bias* refers to differences in item-level responses that occur when items function differently for certain groups of respondents. For example, items are biased when their content is not equally familiar to all groups. Item bias violates the assumption of measurement invariance, which holds that measurement properties should not be affected by the content of the test (Zumbo, 2009).

Regarding the Cognitive Reflection Test, several studies have measured and compared differences across genders, assuming the invariance of the test, although it has never been empirically tested. The aim of Study 1 was to test the equivalence of the CRT-L items across genders by exploring Differential Item Functioning (DIF) within the Item Response Theory (IRT) framework. The aim of the DIF analysis is to ascertain that, after controlling for the underlying construct, the response to an item is related to group membership. If so, the item manifests DIF. For example, if the CRT-L exhibits measurement invariance across genders, a randomly selected woman with a specific level of cognitive reflection and a randomly selected man with the same level of cognitive reflection should have the same chance of giving the correct answer to an item. If this is not the case, DIF is present. In sum, the aim of the current study was to test the invariance property of the CRT-L across genders.

Methods

Participants

The participants were 838 students (52% female; Mean age = 15.3 years; *SD* = 4.03; 55% from secondary school; Mean age = 12.55 years; *SD* = .71; 20% from high school Mean age = 16.27 years; *SD* = 1.5 and 25% attending university; Mean age = 21.12 years; *SD* = 3.5). The adolescents were recruited from secondary and high schools in a suburban area in Italy (Tuscany). A detailed study protocol that explained the goals and methodology of the study was approved by the institutional review boards of each school. Students received an information sheet, which assured them that the data obtained would be handled confidentially and anonymously. All university students were enrolled in the first year of a psychology course at the University of Florence, and were recruited using opportunity sampling from various lectures and seminars. All students participated on a voluntary basis.

In the gender DIF analyses, the male group included 403 students (*Mean age* = 15.03, *SD*= 3.8; *range* = 11.08 to 42), and the female group 435 students (*Mean age* = 15.7, *SD*= 4.2; range = 11.08 to 45)[5].

Materials

The Cognitive Reflection Test-Long (CRT-L, Primi et al., 2016) is an extended version of the CRT (Frederick, 2015) that consists of 6 questions. Although the questions are open-ended, almost all participants produce either the correct response or a typical incorrect (i.e., heuristic) response. That is, the reasoning errors that people make are very systematic. An example item is the following: 'If three elves can wrap three toys in one hour, how many elves are needed to wrap six toys in two hours? [correct answer = 3 elves; heuristic answer = 6 elves]. To be able to produce a correct response, participants have to display an outstanding ability to effectively monitor and correct their impulsive response tendencies. As a result, it is only a small minority of participants who tend to

[5] In the female group the age range was wider than in the male group due to a single participant who was 45 years old. Comparing the means, the difference was significant with a small effect size ($t_{(829)} = 2.449$, $p<.05$, $d = 0.17$).

give correct responses to the tasks. Previous results (Primi et al., 2016) attested that the CRT-L scale is more adequate for younger and less educated samples than the original CRT. Cronbach's alpha[6] in the current study was .80.

Procedure

All students completed the test individually in a self-administered format in the classroom. The task was briefly introduced, and instructions for completion were given. The answers were collected in a paper-and-pencil format. Students were instructed to take as much time as they needed to complete the task. The average administration time was about 5 minutes.

Data Analysis

Preliminarily, we verified the assumption of unidimensionality in each group. In the current study, using the χ^2LD statistic (Chen & Thissen, 1997) we tested the presence of local dependence (LD), i.e., an excess of covariation among item responses that is not accounted for by a unidimensional IRT model. Values of 10 or greater suggest the presence of a multifactorial structure. Otherwise, it is possible to assume that there is a common factor underlying the items.The fit of the IRT model was evaluated using the M_2 statistic and the associated root mean square error of approximation (RMSEA) value. Like other chi-square statistics, the M_2 statistic is generally unrealistic because there will be some error in any strong parametric model (Browne & Cudeck, 1993). Thus, the RMSEA provides a more appropriate metric for model error (Cai, Maydeu-Olivares, Coffman, & Thissen, 2006). RMSEA values of 0.05 or lower indicate good fit. Item parameters were estimated by employing the marginal maximum likelihood estimation method with the expectation–maximization algorithm (Bock, & Aitkin,1981) implemented in IRTPRO (Cai, du Toit, & Thissen, , 2011)., and the item fit under the 2PL model was tested computing the S-χ2 statistics (Orlando & Thissen, 2000). As large samples lead to a greater likelihood of significant

[6] Reliability was .82 for males, and .76 for females.

chi-square differences, the critical value of p=.01, rather than p=.05, was used (Stone & Zhang, 2003).IRT models assume that each examinee responding to a test item possesses some amount of the underlying ability and at each level of ability there will be a certain probability, denoted by $P(\theta)$, to give a correct answer to the item. This approach derives the probability of each response as a function of the latent trait and some item parameters. In the 2PL model the two item parameters are, respectively, item difficulty and item discrimination. The item difficulty parameter (β) or "*location*" measured on the same scale of theta (that conventionally has a mean of zero and SD of 1.0), represents the latent trait level corresponding to a .50 probability of correctly endorsing the item The item discrimination parameter (α) or "*slope"* represents the item's ability to differentiate between people at contiguous levels of the latent trait. This parameter describes how rapidly the probabilities change with trait levels. In order to investigate the invariance property of the items of the scale, analyses of differential item functioning (DIF) across genders were performed, applying an IRT likelihood ratio test approach implemented in the IRTPRO software (Cai, Thissen, & du Toit, 2011).

The DIF detection procedure is based on a nested model comparison approach. For each item, two models are compared, one in which all parameters (discrimination and difficulty) are constrained to be equal across groups, and one with a separate estimation of all parameters. For each item, the fit of a model constraining the item parameters to be equal between the two groups was compared with a model allowing the parameters to be estimated freely in the two groups. This procedure involves comparing differences in log-likelihoods (distributed as chi-square) associated with nested models. Since multiple tests were performed, Bonferroni corrections were used.

Results

Differential Item Functioning (DIF) across gender

The results confirmed that a single factor model adequately represented the structure of the scale for each group, as none of the LD statistics were greater than .10.

The model showed a satisfactory fit (M_2 =16.70, *df* =18, *p*= 0.54; RMSEA =0.0001). Each item had a non-significant S-χ2 value (Table 1), indicating that all items fit under the 2PL model.

Concerning the difficulty parameters (*b*), the results showed that the parameters ranged from -.33 ± .08 to .95 ± .11 logit in the male group and -.39 ± .06 to .87 ± .15 logit in the female group across the continuum of the latent trait. With regard to the discrimination parameters (*a*), according to Baker and Kim (2004), (values 0.01–0.24 are very low, 0.25–0.64 are low, 0.65–1.34 are moderate, 1.35–1.69 are high, and more than 1.7 are very high) all items showed high discrimination levels (*a* values over .1.35) in each group (Table 1).

In the DIF analyses (in which the male group was the reference group), we found gender equivalence for both the discrimination (*a*) and the difficulty (*b*) parameters, after Bonferroni corrections (*p*= .05/ 6 = .008) (Table 1).

Insert Table 1

Discussion

The results attested the measurement equivalence of the scale when administered to male and female students. In other words, the same underlying construct is measured in the two groups. This ensures that the CRT-L can be used to compare males and females, and differences in scores across genders can be interpreted in terms of group differences in the level of the underlying construct.

Study 2

As we described above, Zhang et al. (2016) found that gender differences in cognitive reflection are explained by subjective numeracy. Nevertheless, it is problematic to interpret this result, because the subjective numeracy scale combines the measurement of participants' self-

reported ability and their preference to work with numbers (Fagerlin, Zikmund-Fisher, Ubel, Jankovic, Derry & Smith, 2007). Indeed, it is not clear if subjective numeracy should be considered a measure of numeracy, or if it is more closely related to motivation, emotions, and confidence involving the use of numbers (cf., Liberali, Reyna, Furlan, Stein & Pardo, 2012; Peters & Bjalkebring, 2015). Thus, for a better understanding, it would be desirable to measure numeracy and math-related emotions and attitudes separately.

Regarding the role of cognitive and affective factors in cognitive reflection, Morsanyi et al. (2014) confirmed that both numerical skills and math anxiety independently predicted performance on the CRT. Specifically, they presented a model, where numeracy/school math achievement partially mediated the effect of math anxiety on cognitive reflection (after the effect of test anxiety was controlled). This model was tested both in a sample of female university students, and in a sample of secondary school students. In the secondary school sample, gender was also included in the model as a covariate, and its effect on cognitive reflection was found to be non-significant. That is, gender did not explain additional variance in cognitive reflection, once the effects of math anxiety and math achievement were taken into account. Nevertheless, there was no gender difference in cognitive reflection in this study, and Morsanyi et al. (2014) did not investigate whether the effect of gender on CRT performance was mediated by math skills, math anxiety or both. Thus, the current study sought to replicate the earlier findings regarding the role of math knowledge and math anxiety in cognitive reflection, and additionally test the hypothesis that math knowledge and math anxiety might mediate the effect of gender on cognitive reflection.

Additionally, instead of basic measures of numeracy, which are typically used in studies that explore the links between numerical skills and performance on the CRT, in the current study, we measured more complex mathematical reasoning skills. In order to assess mathematical reasoning abilities comprehensively, we used a combined score of two measures: a measure of probabilistic reasoning skills (including the understanding of basic probabilities presented in text and tables, reasoning about random sequences of events, and the ability to resist some typical fallacies and

biases) and a measure which was developed to assess statistical literacy. Both measures require the application of mathematical/probabilistic reasoning in the context of everyday scenarios. However, the first measure was developed to identify people who struggle with basic probabilistic reasoning, whereas the latter measure is aimed at discriminating between individuals with high levels of statistical reasoning ability. As a result of combining the two scales, we obtained a measure that included items ranging from relatively easy to very difficult, ensuring that participants' mathematical reasoning skills were assessed with good precision.

Methods

Participants

The participants were 181 university students (*Mean age* = 21.23 years, *SD* = 3.67) 33% male (*Mean age* = 21.81 years, *SD* = 4.34) and 66% female (*Mean age* = 20.94 years, *SD* = 3.28)[7] enrolled in the first year of a psychology course at the University of Florence, and were recruited using opportunity sampling from various lectures and seminars. All students participated on a voluntary basis.

Materials

Cognitive reflection was measured by the Italian version of the CRT-L (Primi et al., 2016) as in Study 1. In the present sample, Cronbach's alpha was .76.

Measure of math anxiety. The *Abbreviated Math Anxiety Scale* (AMAS; Hopko, Mahadevan,, Bare & Hunt, 2003; Italian version: Primi, Busdraghi, Tomasetto, Morsanyi, & Chiesi, 2014) measures math anxiety experienced by students in learning and test situations (e.g., "Thinking about an upcoming math test one day before."). Participants have to respond on the basis of how anxious they would feel during the events specified, using a 5-point response scale (ranging from "strongly

[7] There was no age difference across genders (p=.141).

agree" to "strongly disagree"). High scores on the scale indicate high math anxiety. A single composite score was obtained, based on participants' ratings of each statement. In the present sample, Cronbach's alpha was .89.

Measures of mathematical reasoning. The *Berlin Numeracy Test* (BNT, Cokely, Galesic, Schulz, Ghazal, & Garcia-Retamero, 2012) has been developed specifically for educated and highly educated samples (e.g., college students), is composed of 4 questions (with an open-ended question format), and it assesses statistical numeracy and risk literacy. In detail, the contents of the items are about risks, presented in terms of ratio concepts, such as probabilities, proportions, and percentages. (for example "Out of 1,000 people in a small town 500 are members of a choir. Out of these 500 members in the choir 100 are men. Out of the 500 inhabitants that are not in the choir 300 are men. What is the probability that a randomly drawn man is a member of the choir? Please indicate the probability in percent_____ %). A single composite score was computed based on the sum of correct responses. In the present sample, Cronbach's alpha was .50.

The *Probabilistic Reasoning Scale* (PRS, Primi, Morsanyi, Donati, Galli & Chiesi, 2017) has been designed to measure proportional reasoning and basic probabilistic reasoning skills. The scale consists of 16 multiple-choice questions. The items include questions about simple, conditional, and conjunct probabilities, and the numerical data are presented in frequencies or percentages (for example: "A ball was drawn from a bag containing 10 red, 30 white, 20 blue, and 15 yellow balls. What is the probability that it is neither red nor blue?" a) 30/75; b) 10/75; c) 45/75; and "60% of the population in a city are men and 40% are women. 50% of the men and 30% of the women smoke. We select a person from the city at random. What is the probability that this person is a smoker? " a) 42%, b) 50%, c) 85%) A single composite score was computed based on the sum of correct responses. Coefficient alpha for the current sample was .72.

We obtained a measure of *mathematical reasoning* summing the transformed *z* scores of the BNT

and PRS tests. This way, we obtained a measure with a broad range of item difficulty: from easy to very difficult. The reliability of the combined measure was .69.

Procedure

All students completed the measures individually in a self-administered format in the classroom. Each task was briefly introduced, and instructions for completion were given. The answers were collected in a paper-and-pencil format. All measures (the AMAS, PRS, BNT, and CRT-L) were administered in a booklet in a randomized order. There was about half an hour to complete the scales.

Data analysis

To investigate our hypothesis about the relationships between mathematical reasoning, math anxiety, gender and cognitive reflection, we computed Pearson correlations among these variables. To further enhance the understanding of the mechanisms underlying the relationships among these variables, a mediation model was tested. To test our mediation hypothesis concerning the relationship between gender and cognitive reflection through math anxiety and mathematical reasoning, we used the PROCESS macro for SPSS, which allowed us to test a multiple-step multiple mediator model (Hayes, 2009). In this model (Figure 1), path a_1 and path a_2 are, respectively, the regression coefficients estimating math anxiety and mathematical reasoning from gender, and path b_1 and path b_2 are the regression coefficients estimating cognitive reflection from math anxiety and mathematical reasoning, respectively. Path a_3 is the regression coefficient estimating mathematical reasoning from math anxiety. In this model, three specific indirect effects of gender on cognitive reflection can be quantified, i.e., the product of $a_1 \times b_1$, which assesses the indirect effect through math anxiety, and the product of $a_2 \times b_2$, that measures the indirect effect through mathematical reasoning, and the product of $a_1 \times a_3 \times b_2$, which indicates the indirect effect through math anxiety and mathematical reasoning in serial (Figure 1). The sum of the three specific indirect effects corresponds to the total indirect effect (see Brown, 1997).

A bootstrapping procedure (with 10,000 resamples) to estimate 95% bias-corrected confidence intervals (BC 95% CI) was used. A BC 95 % CI that does not include zero provides evidence of a significant indirect effect (Preacher & Hayes, 2008). Bootstrapping is a resampling strategy for estimation and hypothesis testing. With the bootstrapping method, the sample is conceptualized as a pseudo-population that represents the broader population from which the sample was derived, and the sampling distribution of any statistic can be generated by calculating the statistic of interest in multiple resamples from the dataset. The bootstrapping procedure has been suggested as representing the most trustworthy test for assessing the effects of mediation models, overcoming issues associated with inaccurate *p*-values that result from violations of parametric assumptions (Hayes & Scharkow, 2013). Indeed, the bootstrapping procedure is advantageous because it does not impose the assumption of normality of the sampling distribution of indirect effects, and it maintains high power while maintaining adequate control over Type I error rate (Hayes, 2009; MacKinnon, Lockwood, Hoffman, West & Sheets, 2002; MacKinnon, Lockwood & Williams, 2004; Preacher & Hayes, 2008).

Results

Table 2 presents the descriptive statistics and the correlations between the CRT-L and the other measures. As expected, the CRT-L was negatively related to math anxiety and positively related to the BNT and the PRS. The CRT also showed a negative relation with gender: females scored lower on the CRT.

Insert Table 2

To explore the role of math reasoning skills and math anxiety in the gender gap in CRT-L performance, a mediation analysis was conducted with the bootstrapping method with bias-corrected confidence estimates. Results showed that whereas the total effect of gender on cognitive reflection was significant ($p<.001$), once the mediators were entered into the analysis, the direct

effect of gender was no longer significant (p=.245). However, a significant total indirect effect of gender on cognitive reflection was found (point estimate = -0.980, BC 95%CI = -1.363 to -.638). In detail, the results showed a significant indirect effect of gender on cognitive reflection through math anxiety (point estimate = -0.149, BC 95%CI = -.343 to -.025) and mathematical reasoning (point estimate = -0.068, BC 95%CI = -.177 to -.013), and through math anxiety and mathematical reasoning in serial (point estimate = -0.763, BC 95%CI = -1.121 to -.479).

Insert Figure 1

Discussion

The results confirmed that, in line with several previous studies (e.g., Campitelli & Gerrans, 2014; Cueva et. al. 2016; Frederick, 2005; Pennycook et al., 2016; Primi et al., 2016), gender was related to performance on the CRT (i.e., men scored significantly higher than women). However, the gender differences disappeared when numerical skills and math-related anxiety were statistically controlled. Although the CRT is more than just a test of numeracy (e.g., Campitelli & Gerrans, 2014; Liberali et al., 2012), these results suggest that the numerical content of the problems acts as a confounding variable.

Regarding the role of math anxiety in cognitive reflection, a previous study (Morsanyi et al. 2014) showed that highly math anxious individuals responded more quickly (and less accurately) to CRT-problems than participants with low levels of anxiety. This might be interpreted as a strategy to avoid the uncomfortable situation of having to solve numerical problems (see e.g., Ashcraft & Krause, 2007). This tendency might explain why math anxious people are more likely to give incorrect heuristic responses. In other words, the desire to finish the "math task" as soon as possible might prevent them from engaging in more in-depth reasoning and reflection, which is necessary to reach the correct solution. Additionally, previous studies (see Suárez-Pellicioni, Núñez-Peña & Colomé, 2016 for a review) have shown attentional control problems in high math anxious

individuals, especially in inhibitory tasks (Stroop tasks), which could also explain the tendency to rely on salient heuristics. Indeed, several authors have noted that inhibition of the easily available heuristic response is necessary for correct performance on the cognitive reflection test (e.g., Böckenholt, 2012; Campitelli & Gerrans, 2014; Travers et al., 2016). Morsanyi et al. (2014) also explored the possibility that math anxiety affected performance on the CRT through burdening working memory resources. Although both anxiety and working memory load were associated with poorer performance on the CRT, only anxiety (but not working memory load) was associated with faster RTs and lower levels of confidence in responses. For this reason, it was concluded that the effects of anxiety and working memory load were only partially overlapping.

Previous studies have provided evidence for the role of both numeracy and math anxiety/math-related attitudes in performance on the CRT. Some studies have also linked these findings to gender differences in cognitive reflection (Morsanyi et al., 2014; Primi et al., 2016; Zhang et al., 2016). Nevertheless, these studies did not investigate whether the effect of gender on CRT performance was mediated by math skills, math anxiety or both. Additionally, the mediation analysis that was run in the current study also provided some novel information regarding the roles of math reasoning skills and math anxiety. In particular, this model showed that there was a direct link between math anxiety and cognitive reflection but the effect of math anxiety on cognitive reflection was also partially mediated by mathematical reasoning. In sum, the mathematics requirements of the CRT explained gender differences in cognitive reflection – a supposedly domain-general trait-, and when these factors were controlled, men and women did not differ in their level of cognitive reflection. This finding is in line with Thomson and Oppenheimer (2016) who did not find a gender difference in the case of their new version of the CRT (CRT-2), a test that did not require numerical computations.

Ideally, reflectivity should be measured independent of numerical skills and anxiety. Nevertheless, it is not clear if a non-numerical test of cognitive reflection would still be as closely related to decision-making skills as the original CRT or the CRT-L. For example, in Thomson and

Oppenheimer's (2016) study, the correlations between the CRT-2 and various measures of decision-making skills and rational thinking were generally weaker than the correlations between these measures and the original CRT. Indeed, most heuristics and biases tasks and a large proportion of the commonly used decision making competence measures (e.g., the tasks measuring framing effects, the consideration of sample sizes, base rates, intertemporal preferences, risk seeking, etc.) include numerical information. It is also true that many important real-life decisions (including decisions about financial investments, health insurance, pension schemes and medical treatments) require a good level of statistical literacy.

Conclusion

Study 1 attested the measurement equivalence of the CRT-L when administered to male and female participants. Thus, the observed gender differences in cognitive reflection are actual differences and they do not reflect a bias in the measurement process. Although the test measures the same construct in both males and females, the results of Study 2 suggest that numerical skills and math-related feelings, such as math anxiety, are strongly correlated with performance on the CRT. Given that men often outperform women on math reasoning tasks and they show less anxiety, this finding is problematic for studies that seek to measure reflectivity independently of quantitative skills. One solution to this problem could be to measure cognitive reflection independent of math skills (e.g., Thomson & Oppenheimer, 2016). Nevertheless, this line of work needs further validation, and it would also be advantageous to develop additional tasks for this purpose, which cover a broader range of difficulty levels. Another possible solution to this issue could be to use math skills as a covariate when looking at the relations between the CRT and other measures in order to be able to distinguish between the effects of reflection and numeracy. Nevertheless, in this case, the results would strongly depend on the choice of numeracy measure. Future studies should aim to explore further the role of reflectivity and numerical skills in distinguishing between competent and incompetent decision makers. Additionally, these studies

could also confirm whether gender differences disappear when reflectivity is measured independently of quantitative skills.

References

Albaity, M., Rahman, M., & Shahidul, I. (2014). Cognitive reflection test and behavioral biases in Malaysia. *Judgment & Decision Making*, *9*(2), 149-151.

Ashcraft, M. H., & Krause, J. A. (2007). Working memory, math performance, and math anxiety. *Psychonomic Bulletin & Review*, *14*, 243-248.

Ashcraft, M. H., & Ridley, K. S. (2005). Math anxiety and its cognitive consequences: A tutorial review. In J. I. D. Campbell (Ed.), *Handbook of Mathematical Cognition* (pp. 315-327). New York: Psychology Press

Baker, F. B. (2001). *The basics of item response theory*. For full text: http://ericae. net/irt/baker.

Baker, F. B., & Kim, S. H., (eds.). (2004). *Item Response Theory: Parameter Estimation Techniques*, 2nd Edn. New York, NY: Marcel Dekker

Baloglu, M., & Kocak, R. (2006). A multivariate investigation of the differences in mathematics anxiety. *Personality and Individual Differences*, *40*(7), 1325-1335.

Beasley, T. M., Long, J. D., & Natali, M. (2001). A confirmatory factor analysis of the mathematics anxiety scale for children. *Measurement and Evaluation in Counseling and Development*, *34*(1), 14-26.

Beilock, S. L. (2008). Math performance in stressful situations. *Current Directions in Psychological Science, 17*(5), 339–334.

Benbow, C. P., Lubinski, D., Shea, D. L., & Eftekhari-Sanjani, H. (2000). Sex differences in mathematical reasoning ability at age 13: Their status 20 years later. *Psychological Science*, *11*(6), 474-480.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 443-459.

Böckenholt, U. (2012). The cognitive-miser response model: Testing for intuitive and deliberate reasoning. *Psychometrika, 77*, 388–399.

Brañas-Garza, P., Kujal, P., & Lenkei, B. (2015). Cognitive Reflection Test: Whom, How, When. Retrieved, 25.01.17, from https://mpra.ub.uni-muenchen.de/68049/

Brown, R. L. (1997). Assessing specific mediational effects in complex theoretical models. Structural Equation Modeling, 4, 142156.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. *Sage focus editions*, *154*, 136-136.

Bursal, M., & Paznokas, L. (2006). Mathematics anxiety and preservice elementary teachers' confidence to teach mathematics and science. *School Science and Mathematics*, *106*(4), 173-180.

Cai, L., Du Toit, S. H. C., & Thissen, D. (2011). IRTPRO: Flexible, multidimensional, multiple categorical IRT modeling [Computer software]. *Chicago, IL: Scientific Software International*.

Cai, L., Maydeu-Olivares, A., Coffman, D. L., & Thissen, D. (2006). Limited-information goodness-of-fit testing of item response theory models for sparse 2P tables. *British Journal of Mathematical and Statistical Psychology*, *59*(1), 173-194.

Campitelli, G., & Gerrans, P. (2014). Does the cognitive reflection test measure cognitive reflection? A mathematical modeling approach. *Memory & Cognition*, *42*, 434-447.

Chen, W. H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, *22*(3), 265-289.

Chipman, Susan F., David H. Krantz, and Rae Silver. "Mathematics anxiety and science careers among able college women." *Psychological science* 3.5 (1992): 292-296.

Cokely, E. T., Galesic, M., Schulz, E., Ghazal, S., & Garcia-Retamero, R. (2012). Measuring risk literacy: The Berlin numeracy test. *Judgment and Decision Making*, *7*(1), 25-47.

Cokely, E. T. Y., & Kelly, C. M. (2009). Cognitive abilities and superior decision making under risk: A protocol analysis and process model evaluation. *Judgment and Decision Making*, *4*, 20-33.

Cueva, C., Iturbe-Ormaetxe, I., Mata-Pérez, E., Ponti, G., Sartarelli, M., Yu, H., Zhukova, V., 2016. Cognitive (ir)reflection: New experimental evidence. *J. Behav. Exp. Econ.* 64, 81–93.

Devine, A., Fawcett, K., Szűcs, D., & Dowker, A. (2012). Gender differences in mathematics anxiety and the relation to mathematics performance while controlling for test anxiety. *Behavioral and Brain Functions*, *8*(1), 33.

Else-Quest, N. M., Hyde, J. S., & Linn, M. C. (2010). Cross-national patterns of gender differences in mathematics: A meta-analysis. *Psychological Bulletin*, *136*(1), 103-127.

Fagerlin, A., Zikmund-Fisher, B. J., Ubel, P. A., Jankovic, A., Derry, H. A., & Smith, D. M. (2007). Measuring numeracy without a math test: development of the Subjective Numeracy Scale. *Medical Decision Making*, *27*(5), 672-680.

Fernbach, P. M., Rogers, T., Fox, C. R., & Sloman, S. A. (2013). Political extremism is supported by an illusion of understanding. *Psychological science*, *24*(6), 939-946.

Fernbach, P. M., Sloman, S. A., Louis, R. S., & Shube, J. N. (2013). Explanation fiends and foes: How mechanistic detail determines understanding and preference. *Journal of Consumer Research*, *39*(5), 1115-1131.

Finucane, M. L., & Gullion, C. M. (2010). Developing a tool for measuring the decision-making competence of older adults. *Psychology and Aging*, *25*(2), 271.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, *19*, 25-42.

Hayes, A. F. (2009). Beyond Baron and Kenny: Statistical mediation analysis in the new millennium. Communication monographs, 76(4), 408-420.

Hayes, A. F., & Scharkow, M. (2013). The relative trustworthiness of inferential tests of the indirect effect in statistical mediation analysis does method really matter?. *Psychological science*, *24*(10), 1918-1927.

Hembree, R. (1990). The nature, effects, and relief of mathematics anxiety. *Journal for Research in Mathematics Education*, 33-46

Hill, F., Mammarella, I. C., Devine, A., Caviola, S., Passolunghi, M. C., & Szűcs, D. (2016). Maths anxiety in primary and secondary school students: Gender differences, developmental changes and anxiety specificity. *Learning and Individual Differences*. doi:10.1016/j.lindif.2016.02.006

Hyde, J. S., Fennema, E., & Lamon, S. J. (1990). Gender differences in mathematics performance: a meta-analysis.

Hyde, J. S., Fennema, E., Ryan, M., Frost, L. A., & Hopp, C. (1990). Gender comparisons of mathematics attitudes and affect. *Psychology of Women Quarterly, 14*(3), 299-324.

Hyde, J. S., Lindberg, S. M., Linn, M. C., Ellis, A. B., & Williams, C. C. (2008). Gender similarities characterize math performance. *Science, 321*, 494-495.

Ho, H. Z., Senturk, D., Lam, A. G., Zimmer, J. M., Hong, S., Okamoto, Y., ... & Wang, C. P. (2000). The affective and cognitive dimensions of math anxiety: A cross-national study. *Journal for Research in Mathematics Education, 31*(3), 362-379

Hopko, D. R., Mahadevan, R., Bare, R. L., & Hunt, M. K. (2003). The abbreviated math anxiety scale (AMAS) construction, validity, and reliability. *Assessment, 10*(2), 178-182.

Jain, S., & Dowson, M. (2009). Mathematics anxiety as a function of multidimensional self-regulation and self-efficacy. *Contemporary Educational Psychology*, *34*(3), 240-249

Jameson, M. M., & Fusco, B. R. (2014). Math anxiety, math self-concept, and math self-efficacy in adult learners compared to traditional undergraduate students. *Adult Education Quarterly*, 64(4), 306-322.

Koehler, D. J., & James, G. (2010). Probability matching and strategy availability. *Memory & Cognition*, *38*(6), 667-676.

Gómez-Chacón, I. M., García-Madruga, J. A., Vila, J. Ó., Elosúa, M. R., & Rodríguez, R. (2014). The dual processes hypothesis in mathematics performance: Beliefs, cognitive reflection, working memory and reasoning. *Learning and Individual Differences*, *29*, 67-73.

Liberali, J. M., Reyna, V. F., Furlan, S., Stein, L. M., & Pardo, S. T. (2012). Individual differences in numeracy and cognitive reflection, with implications for biases and fallacies in probability judgment. *Journal of Behavioral Decision Making*, *25*(4), 361-381.

McGraw, R., Lubienski, S. T., & Strutchens, M. E. (2006). A closer look at gender in NAEP mathematics achievement and affect data: Intersections with achievement, race/ethnicity, and socioeconomic status. *Journal for Research in Mathematics Education, 37*(2), 129-150

McMullan, M., Jones, R., & Lea, S. (2010). Patient safety: numerical skills and drug calculation abilities of nursing students and registered nurses. *Journal of Advanced Nursing*, *66*(4), 891-899

MacKinnon, D. P., Lockwood, C. M., & Williams, J. (2004). Confidence limits for the indirect effect: Distribution of the product and resampling methods. *Multivariate behavioral research*, *39*(1), 99-128.

MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G., & Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychological methods*, *7*(1), 83.

Maloney, E. A., & Beilock, S. L. (2012). Math anxiety: Who has it, why it develops, and how to guard against it. *Trends in Cognitive Sciences*, *16*(8), 404-406

Mau, W. C., & Lynn, R. (2000). Gender differences in homework and test scores in mathematics, reading and science at tenth and twelfth grade. *Psychology, Evolution and Gender*, *2*(2), 119-125.

Morsanyi, K., Busdraghi, C., & Primi, C. (2014). Mathematical anxiety is linked to reduced cognitive reflection: a potential road from discomfort in the mathematics classroom to susceptibility to biases. *Behavioral and Brain Functions*, *10*(1), 31.

Morsanyi, K., McCormack, T. & O'Mahony, E. (2018). The link between deductive reasoning and mathematics. *Thinking and Reasoning (in press)* DOI:10.1080/13546783.2017.1384760.

Morsanyi, K., Primi, C., Handley, S.J., Chiesi, F. & Galli, S. (2012). Are systemizing and autistic traits related to talent and interest in mathematics and engineering? Testing some of the central claims of the empathizing-systemizing theory. *British Journal of Psychology, 103,* 472-496.

Mata, A., Fiedler, K., Ferreira, M. B., & Almeida, T. (2013). Reasoning about others' reasoning. *Journal of Experimental Social Psychology*, *49*(3), 486-491.

OECD (2013). *PISA 2012 Assessment and analytical framework: mathematics, reading, science, problem solving and financial literacy*. OECD Publishing

OECD (2016), *PISA 2015 Results (Volume 1): Excellence and Equity in Education.* OECD Publishing, Paris http://www.oecd-ilibrary.org/education/pisa-2015-results-volume-i_9789264266490-en

Oechssler, J., Roider, A., & Schmitz, P. W. (2009). Cognitive abilities and behavioral biases. *Journal of Economic Behavior & Organization*, *72*(1), 147-152

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, *24*(1), 50-64.

Pennycook, G., Cheyne, J. A., Seli, P., Koehler, D. J., & Fugelsang, J. A. (2012). Analytic cognitive style predicts religious and paranormal belief. *Cognition*, *123*(3), 335-346.

Pennycook, G., Cheyne, J. A., Koehler, D. J., & Fugelsang, J. A. (2016). Is the cognitive reflection test a measure of both reflection and intuition?. *Behavior Research Methods*, *48*(1), 341-348.

Peters, E., & Bjalkebring, P. (2015). Multiple numeric competencies: When a number is not just a number. *Journal of personality and social psychology*, *108*(5), 802.

Pozehl, B. J. (1996). Mathematical calculation ability and mathematical anxiety of baccalaureate nursing students. *Journal of Nursing Education*, *35*(1), 37-39

Preacher, K. J., & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior research methods*, *40*(3), 879-891.

Primi, C., Busdraghi, C., Tomasetto, C., Morsanyi, K., & Chiesi, F. (2014). Measuring math anxiety in Italian college and high school students: validity, reliability and gender invariance of the Abbreviated Math Anxiety Scale (AMAS). *Learning and Individual Differences*, *34*, 51-56.

Primi, C., Morsanyi, K., Chiesi, F., Donati, M. A., & Hamilton, J. (2016). The development and testing of a new version of the cognitive reflection test applying item response theory (IRT). *Journal of Behavioral Decision Making*, *29*, 453-469.

Primi, C., Morsanyi, K., Donati,M., Galli, S., & Chiesi, F. (2017). Measuring Probabilistic Reasoning: The Construction of a New Scale Applying Item Response Theory. *Journal of Behavioral Decision Making* doi: 10.1002/bdm.2011

Ring, P., Neyse, L., David-Barett, T., & Schmidt, U. (2016). Gender Differences in Performance Predictions: Evidence from the Cognitive Reflection Test. *Frontiers in Psychology*, *7*.

Rubinsten, O., Bialik, N., & Solar, Y. (2012). Exploring the relationship between math anxiety and gender through implicit measurement. *Frontiers in Human Neuroscience*, *6*, 279.

Shenhav, A., Rand, D. G., & Greene, J. D. (2012). Divine intuition: cognitive style influences belief in God. *Journal of Experimental Psychology: General*, *141*(3), 423.

Sinayev, A., & Peters, E. (2015). Cognitive reflection vs. calculation in decision making. *Frontiers in Psychology*, *6*.

Stone, C. A., & Zhang, B. (2003). Assessing goodness of fit of item response theory models: A comparison of traditional and alternative procedures. *Journal of Educational Measurement*, *40*(4), 331-352.

Suárez-Pellicioni, M., Núñez-Peña, M., & Colomé, À. (2016). Math anxiety: A review of its cognitive consequences, psychophysiological correlates, and brain bases. *Cognitive, affective & behavioral neuroscience*, *16*(1).

Swars, S. L., Daane, C. J., & Giesen, J. (2006). Mathematics anxiety and mathematics teacher efficacy: What is the relationship in elementary preservice teachers? *School Science and Mathematics, 106*(7), 306-315.

Thomson, K. S., & Oppenheimer, D. M. (2016). Investigating an alternate form of the cognitive reflection test. *Judgment and Decision Making*, *11*(1), 99-113.

Toplak, M. E., West, R. F., & Stanovich, K. E. (2011). The Cognitive Reflection Test as a predictor of performance on heuristics-and-biases tasks. *Memory & Cognition*, *39*(7), 1275.

Toplak, M. E., West, R. F., & Stanovich, K. E. (2014). Assessing miserly information processing: An expansion of the Cognitive Reflection Test. *Thinking & Reasoning*, *20*(2), 147-168.

Van de Vijver, F. J., & Poortinga, Y. H. (1997). Towards an integrated analysis of bias in cross-cultural assessment. *European journal of psychological assessment*, *13*(1), 29.

Zhang, D. C., Highhouse, S., & Rada, T. B. (2016). Explaining sex differences on the Cognitive Reflection Test. *Personality and Individual Differences*, *101*, 425-427.

Zumbo, B. D. (2009). Validity as contextualized and pragmatic explanation, and its implications for validation practice. *The concept of validity: Revisions, new directions and applications*, 65-82.

Waiyavutti, C., Johnson, W., & Deary, I. J. (2012). Do personality scale items function differently in people with high and low IQ?. *Psychological assessment*, *24*(3), 545.

Weller, J. A., Dieckmann, N. F., Tusler, M., Mertz, C. K., Burns, W. J., & Peters, E. (2013). Development and testing of an abbreviated numeracy scale: A Rasch analysis approach. *Journal of Behavioral Decision Making*, *26*(2), 198-212.

Wu, S. S., Barth, M., Amin, H., Malcarne, V., & Menon, V. (2012). Math anxiety in second and third graders and its relation to mathematics achievement. *Frontiers in Psychology*, *3*, 162-162.

Table 1.

Fit statistics, parameters for each item of the CRT-L for gender groups, and DIF analysis of discrimination and difficulty parameters across genders.

	Males				*Females*				*a DIF*			*b DIF*		
Item	*S-χ^2* (*df*)	*p*	*a* (*SE*)	*b* (*SE*)	*S-χ^2* (*df*)	*p*	*a* (*SE*)	*b* (*SE*)	χ^2	*df*	*p*	χ^2	*df*	*p*
1	.99 (4)	.912	2.38 (.37)	.70 (.09)	.94 (4)	.919	2.29 (.37)	.71 (.10)	0.0	1	.87	0.0	1	.97
2	6.13 (4)	.189	1.99 (.29)	.59 (.10)	6.05 (4)	.195	3.98 (.80)	.48 (.06)	5.5	1	.02	0.1	1	.80
3	.69 (4)	.952	3.47 (.64)	.34 (.07)	3.93 (4)	.417	3.32 (.57)	.39 (.06)	0.0	1	.86	0.2	1	.63
4	6.95 (2)	.031	3.41 (.68)	-.33 (.08)	.57 (2)	.751	3.11 (1.20)	-.39 (.06)	0.1	1	.73	0.2	1	.66
5	1.73 (4)	.786	2.40 (.38)	.95 (.11)	1.72 (4)	.787	2.80 (.50)	.77 (.09)	0.4	1	.53	1.1	1	.29
6	.47 (4)	.978	2.43 (.37)	.54 (.08)	1.73 (4)	.785	1.49 (.26)	.87 (.15)	4.3	1	.04	1.8	1	.19

Note. *S-χ^2* statistics, *df* = degrees of freedom, (Degrees of freedom are equal to the difference between the number of theta levels and the number of the model parameters). Depending on the sample characteristics, the number of theta levels might change from one item to the other because the number of cases for each theta level might be too small for some items. If this happens, some consecutive levels are collapsed in order to reach an adequate number of cases for all the levels, which allows more stable estimates).; parameters: *a* = discrimination, *b* = difficulty., *SE* = standard error, *DIF* = Differential Item Functioning (Due to the large sample size α was fixed at .01).

Table 2.

Correlations between the CRT-L, gender, the BNT, probabilistic reasoning ability and maths anxiety.

	CRT-L	BNT	PRS	Maths Reasoning	AMAS	Gender
1 CRT-L						
2 BNT	.42***					
3 PRS	.55***	.38***.				
4 Maths Reasoning	.59***	.83***	:83***			
4 AMAS	-.39***	-.23**	-.24**	-.29***		
5 Gender	-.33***	-.41***	-.32***	-.44***	.16*	
M	2.82	1.10	12.51	0	26.36	
SD	1.86	1.03	2.42	1	7.92	

Note. Males coded as 1; females coded as 2. CRT-L = Cognitive Reflection Test – Long, , BNT = Berlin Numeracy Test; PRS = Probabilistic Reasoning Scale, Maths Reasoning (the combined z scores of the BNS and PRS) AMAS = Abbreviated Math Anxiety Scale.
****p* < .001, ***p* < .01, **p*<.05

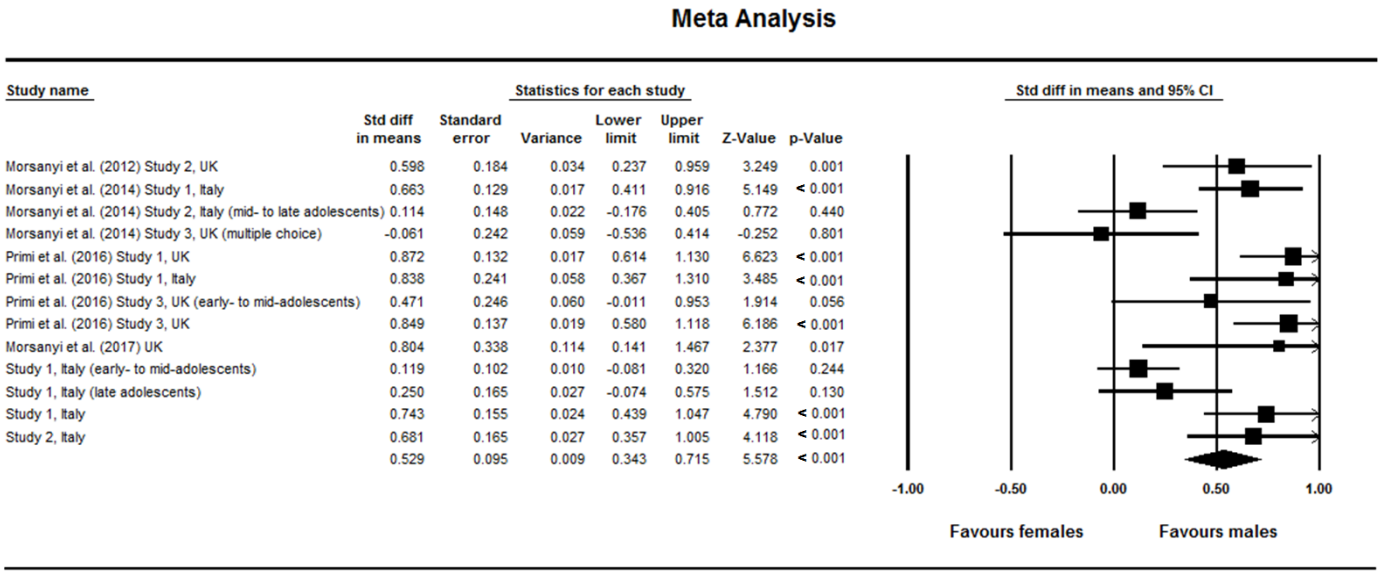

Figure 1 Meta-analysis of gender differences in cognitive reflection based on studies conducted by our research group with British and Italian participants.

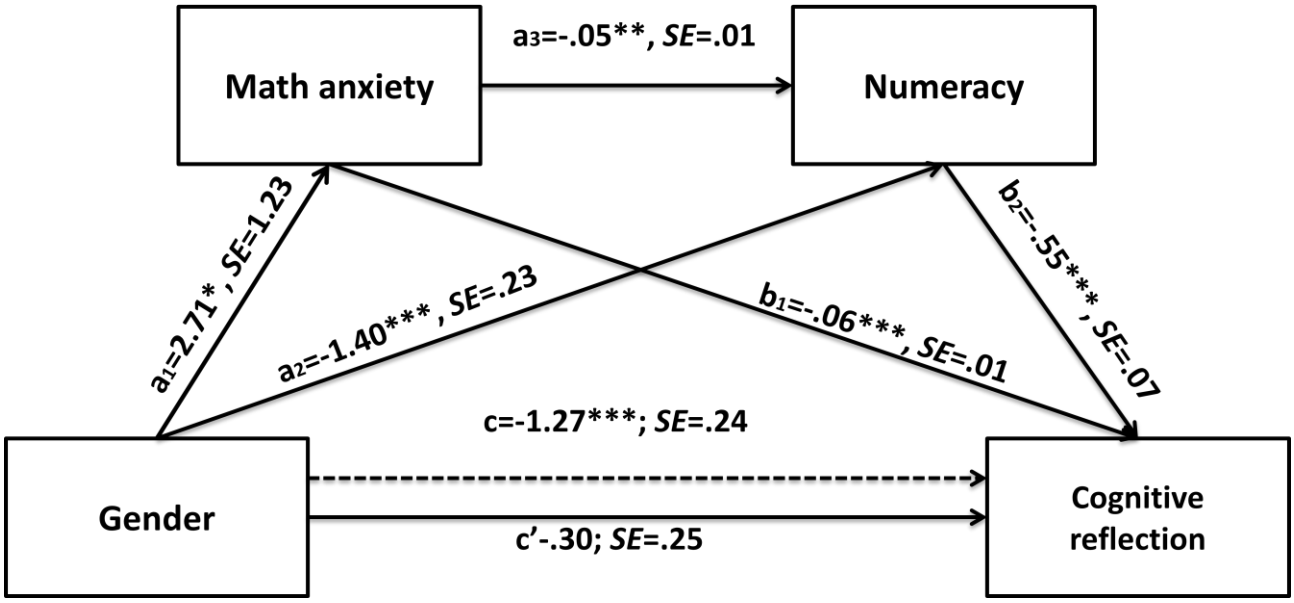

Figure 2. Multiple-step multiple mediator model with gender as independent variable, math anxiety and mathematical reasoning as simultaneous mediators, and cognitive reflection as a dependent variable.

NOTE: Path values represent unstandardized ordinary least squares (OLS) regression coefficients. Dotted line (c) represents the total effect of gender on cognitive reflection, i.e., the effect prior to the inclusion of the mediating variables. The c' value represents the direct effect of gender on cognitive reflection, i.e. the effect of gender on cognitive reflection after the mediators are included. *p < .05; **p < .01; ***p < .001.